

(19) 日本国特許庁 (J P)

## (12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-167001

(P2001-167001A)

(43) 公開日 平成13年6月22日 (2001.6.22)

(51) Int.Cl. <sup>7</sup>	識別記号	F I	テーマコード* (参考)
G 0 6 F 12/16	3 1 0	G 0 6 F 12/16	3 1 0 Q
	3 2 0		3 1 0 J
			3 2 0 F
			3 2 0 L

審査請求 未請求 請求項の数 1 O L (全 11 頁)

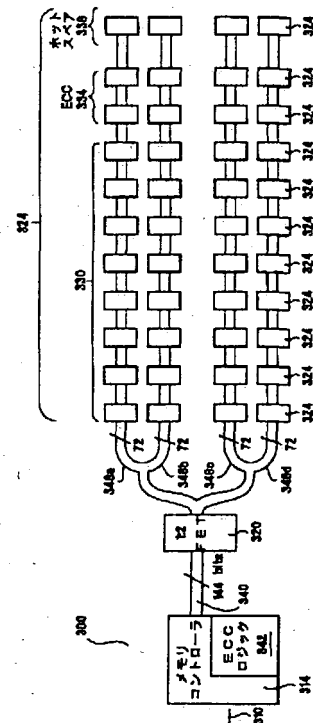
(21) 出願番号	特願2000-312778 (P2000-312778)	(71) 出願人	398038580 ヒューレット・パカード・カンパニー HEWLETT-PACKARD COM PANY アメリカ合衆国カリフォルニア州パロアル ト ハノーバー・ストリート 3000
(22) 出願日	平成12年10月13日 (2000. 10. 13)	(72) 発明者	マイケル・ビー・レイハム アメリカ合衆国95033カリフォルニア州ロ ス・ガトス、クヌース・ロード 18219
(31) 優先権主張番号	0 9 / 4 3 0 3 6 3	(72) 発明者	ジェームス・ジー・マティオス アメリカ合衆国95070カリフォルニア州サ ラトガ、ウッドサイド・ドライブ 12541
(32) 優先日	平成11年10月28日 (1999. 10. 28)	(74) 代理人	100081721 弁理士 岡田 次生
(33) 優先権主張国	米国 (US)		

(54) 【発明の名称】 自己回復するメモリ構成

(57) 【要約】

【課題】 複雑なデータ修復方法を必要とせず、複数チャネル上の複数エラーに対するエラー検出および訂正を行うと共に、障害の発生したメモリのホットスワッピングもサポートする、メモリシステム構成を提供する。

【解決手段】 本メモリ構成には、メモリコントローラに電気的に結合されているCPUバスと、スイッチング手段に接続されたメモリコントローラと、複数のメモリモジュールに電気的に接続されたスイッチを含む。複数のメモリモジュールは、少なくとも一つのデータ用のメモリモジュールと、少なくとも一つのデータ訂正用のメモリモジュールと、少なくとも一つのホットスペアモジュールとを含む。データ用メモリモジュールと、エラー訂正用メモリモジュールと、ホットスペアメモリモジュールとは並列に接続されている。



# 【特許請求の範囲】

【請求項1】 メモリコントローラに電氣的に結合されたCPUバスと、

前記メモリコントローラに電氣的に結合されたデータバスと、

少なくとも1つのデータ記憶用メモリモジュールと、

少なくとも1つのスペアメモリモジュールと、を具備し、

前記少なくとも1つのデータ記憶用メモリモジュールおよび前記少なくとも一つのスペアメモリモジュールは、前記データバスに並列に電氣的に結合されており、前記少なくとも1つのデータ記憶用メモリモジュールで発生する訂正可能なエラーに回答して、前記少なくとも1つのデータ記憶用メモリモジュールからのデータが前記スペアメモリモジュールにマップされるメモリ構成。

# 【発明の詳細な説明】

## 【0001】

【発明の属する技術分野】本発明は、メモリ障害後に自己回復することができるサーバ用のメモリシステムに関する。

## 【0002】

【従来の技術】利用可能なサーバメモリシステムのサイズは、時と共に増大し続けており、現サーバメモリシステムは、64ギガバイト以上であることも多い。メモリシステムのサイズが増大するに従い、メモリビットの障害、すなわちメモリシステムに障害が発生する可能性も増加している。メモリシステム障害は、一時的エラーまたは永続的エラーの両方で起こり得る。一時的エラーは、一般に、アルファ粒子によって起こる。永続的エラーには、一般に、メモリセルウォール障害、行または列デコード障害、状態機械障害、またはデュアルインラインメモリモジュール（以下、「DIMM」とする）の印刷回路基板とシステム印刷回路基板間の機械的インタフェースの故障といった他の破局故障などがある。

【0003】メモリシステムの障害を防止するために、種々の形態のメモリ検出および訂正プロセスが開発されている。一般的に使用されているシステムのうちの1つでは、エラーを検出するパリティビットを使用する。データを受信すると、予期される値に対してデータのパリティがチェックされる。データが予期されるパリティ値（奇数または偶数）と一致しない場合は、エラーが発生したと判断される。この方法は単一のビットエラーを判断するには機能するが、複数のビットエラーを判断するには必ずしも適切に作用するものではない。更に、最も簡単なパリティシステムは、データエラーを訂正する仕組みを有していない。

【0004】Gaskinsらに対する米国特許第5,463,643号は、複数エラーを検出できるパリティシステムを提供する。Gaskinsらの構成は、データエラーを訂正する機構も含む。RAMBUSメモリシステムに動作するよう

設計されている。RAMBUS仕様には、16ビットデータバスとパリティビットと種々の制御信号、電源信号および接地信号とを含むメモリチャンネル仕様が含まれる。

【0005】図1は、直列に接続された複数のメモリモジュールを有する一般的なRAMBUSメモリ構成100の高レベル部分ブロック図を示す。図1には、メモリチャンネル112およびCPUバス114に電氣的に接続されたメモリI/Oコントローラ110が示されている。メモリチャンネル112は、RAMBUS構成仕様に従って直列に接続された複数のメモリモジュール116a, 116b, ..., 116nに電氣的に接続されている。

【0006】Gaskinsらによって述べられているRAMBUS仕様に関し、起こり得る問題の1つとして、パリティチェックがメモリエラーの検出にのみ供されており、エラー訂正用には限定的な提供しかされていない、ということがある。Gaskinsらの構成はデータエラーの訂正はできるが、そのエラー訂正能力は限定的である。図2を参照すると、4つのデータチャンネル210a~210dおよび1つのパリティチャンネル210eを有する、Gaskinsらのデータチャンネルおよびパリティチャンネルの高レベル部分ブロック図が示されている。Gaskinsらの構成は単一チャンネル上の単一または複数のエラーを訂正することができるが、出願人は、この構成では複数チャンネル上で2つ以上のエラーを検出することができないと考えている。例えば、チャンネル210aで完全な（すべてのビット）障害が発生した場合は、データを再構成することができる。しかしながら、異なるチャンネル上で2つのビット障害が発生した（例えば、チャンネル210a上で1つのビット障害が発生し、チャンネル210b上で1つのビット障害が発生する）場合、エラーの検出はできず、データを再構成できない。

【0007】Gaskinsらの構成における更なる問題は、メモリI/Oコントローラ上のI/Oピン数に関連する。メモリ記憶要件の増加につれ、メモリコントローラに接続されたメモリモジュールの数も増加する。直列アーキテクチャ構成では、メモリチャンネルの数を1つ増やすと、メモリコントローラに直接接続されたメモリチャンネルの数も1つ増える。各メモリチャンネルを追加するごとに入力ピンを追加する必要があり、システムピン数が増加するので、これは問題となる。この問題は、特に大容量のメモリを有するメモリ記憶システムにおいて顕著となる。

【0008】通常使用される他のエラー検出プロセスは、エラー訂正またはエラーチェックおよびエラー訂正コード（以下、ECCと呼称する）である。一般に、ECCは、CRC（巡回冗長検査）アルゴリズムに基づいている。ECCコードは、原則としてCRCコードであり、その冗長度は十分に高いので、エラーが破局的でなければ、オリジナルデータを復元することができる。C

RCアルゴリズムは、データを受信したときに完全なデータシーケンス（データフィールドの終端に付加されたCRCビットを含む）がCRCチェッカによって読み出されるように動作する。完全なデータシーケンスは、CRC多項式によって余りなく割切れなければならない。完全なデータシーケンスがCRC多項式によって割切れない場合、エラーが発生していると判断される。

#### 【0009】

【発明が解決しようとする課題】パリティに基づく従来からのエラー訂正プロセスと異なり、ECCコードに基づくシステムは、一般に、複数のビットエラーを検出するために使用できる。例えば、単一ビット訂正を持つECCメモリシステムは、一般的に、二重ビットエラーを検出し単一ビットエラーを訂正することができる。4または8ビットエラー訂正を持つECCメモリは、一般的に、4ビットまたは8ビットエラーを検出し訂正することができる。従って、 $\times 4$ または $\times 8$ 構成で編成されたSDRAMチップ全体の障害により、システムに障害が発生することはない。ECCシステムは容易に複数エラービットの検出を行うが、従来からのECCシステムの問題は、一般に、それらが訂正不可能なエラーを報告する場合にシステムを停止させてしまうということである。このため、最初の停止をしないで、ECCメモリシステム中の障害部分を置き換えて、システムの障害耐力（immunity：イミュニティ）を回復することはできない。

【0010】ECCおよびパリティプロセスは、一般に、週7日、24時間の動作を要求される、半導体メモリを用いるコンピュータシステムにおいて使用されている。コンピュータシステムが半導体メモリに要求されている速度を必要としない場合か、または半導体メモリを用いた、対コスト上有効でさらに大容量の記憶装置を利用することができない場合は、ディスクドライブメモリが使用される。ディスクドライブメモリによってサポートされるコンピュータシステムは、一般に、ハードウェアによって複数のディスクドライブを互いにリンクさせて、廉価ディスク冗長アレイ（redundant array of inexpensive disks）すなわちRAIDとして周知の、ドライブアレイを形成する。アレイ中のドライブは、互いに協調しており、データはそれらの間に特別な方法で割り振られる。ディスクドライブおよびディスクドライブを読み出す機械的インタフェースは、半導体メモリより信頼性が低いので、ディスクドライブシステムにおける永続的または一時的なデータ障害のデータ回復プロセスは、一般に、より多くの冗長性とより複雑なデータ回復手順を有している。

【0011】従来のRAIDシステムにおいては、データはドライブ間でビットまたはバイトレベルで分割される。例えば、4ドライブシステムにおいて、各バイトの2ビットは第1のハードディスクからもたらされ、次の

2ビットは第2のハードディスクからもたらされる、というようにである。従って、4つのドライブは、ドライブを直列にしたときより4倍の速度で単一バイトのデータストリームを出力する。すなわち、1つのドライブが2ビットを転送するために必要な時間と同じ時間しかかからないで、1バイトで構成された情報のすべてが転送される。いくつかのドライブ間でデータを分割するこの技術は、データストライピングと呼ばれ、ドライブ毎の実際のブロックサイズは1000バイト程度の大きにすることができる。RAIDメモリにより、各ディスクを独立して動作させる場合よりも信頼性が向上し、エラーに対してより大きな抵抗力を持たせることができる。信頼性および障害耐性（fault-tolerance）の向上は、ミラリングおよびパリティの実施を含む種々の冗長方法によって達成される。

【0012】複雑なデータ回復方法論を必要とせず、複数チャンネル上の複数エラーに対する自動的なエラー検出および訂正機能を備えると共に、障害メモリのホットスワッピングもサポートする、メモリシステム構成が必要とされている。

#### 【0013】

【課題を解決するための手段】本発明は、メモリ障害の後に自己回復（self heal）し、システムのシャットダウンまたはオペレータの介入無しで障害耐力を正常レベルに回復できる、サーバ用の自己回復メモリシステムを提供する。このメモリシステムは、「自己リカバリ」、つまり、アルファ粒子またはSDRAMのビットセルウォール障害によってより頻発する単一ビット障害と同様に、DIMMにおけるSDRAM障害、DIMMコネクタ障害の回復が可能である。

【0014】本発明の自己回復メモリ構成は、SDRAM（SDRまたはDDR）のような並列メモリ技術またはRAMBUSのような直列メモリ技術と共に使用することができる。好ましい実施の形態では、メモリ構成は、メモリコントローラに電気的に結合されたCPUバスと、スイッチング手段に接続されたメモリコントローラと、複数のメモリモジュールに電気的に接続されたスイッチング手段と、を備えている。ここで、複数のメモリモジュールは、複数のデータメモリモジュールと、複数のECCメモリモジュールと、ホットスเปアメモリモジュールと、を含み、これら複数のデータメモリモジュールとECCメモリモジュールとホットスเปアメモリモジュールとは、並列に接続されている。1または4ビットの訂正可能なハードエラーに対して、エラーが発生したデータメモリモジュールに格納されたデータは、ホットスเปアメモリモジュールにマップされる。複数のデータ、ECCおよびスぺアメモリモジュールの並列構成により、CPUのローディングの低減が容易になり、また必要な場合には、障害が発生したメモリコンポーネントのホットスワッピングが容易となる。

【0015】好ましい実施の形態は、ホットスเปアメモリモジュールを含む。ECC DIMMおよびホットスペアを使用するSDRAMメモリシステムは、SDRAMの障害またはDIMMコネクタの障害を原因とする、障害が発生したDIMMからのデータの再構築を提供する。そして、障害の発生したDIMMを、オンラインで、あるいはそのシステムがホットスワップをサポートしていない場合は年一度のダウンタイムのうちの数分間に、置換することができる。その後、置換されたDIMMメモリモジュールは、メモリシステムのホットスペアとなる。

【0016】本発明の特性および利点は、明細書の残りの部分および添付図面とに関連して更に理解されよう。

【0017】

【発明の実施の形態】図3は、本発明による自己回復並列メモリ構成300の部分ブロック図を示す。好ましい実施の形態では、メモリ構成は、メモリコントローラ314に電気的に結合されたCPUバス310と、スイッチング手段320に電気的に結合されたメモリコントローラ314と、を備えており、スイッチング手段320は、複数のメモリモジュール324に電気的に接続されている。複数のメモリモジュール324は、好ましくは、複数のデータメモリモジュール330と、複数のECCメモリモジュール334と、複数のスペアメモリモジュール338と、を有している。好ましい実施の形態では、データメモリモジュール330とECCメモリモジュール334とホットスペアメモリモジュール338とは、並列に接続されている。

【0018】図3に示す実施の形態では、データメモリ構成は、144ビットのデータメモリチャンネル340と、36のメモリモジュールと、を有する自己回復DDRシステムであり、好ましくは、データメモリモジュール330用に32のDDR DIMM、ECCメモリモジュール334用に8つのDIMM、および4つのDDRホットスペアDIMM338を備えている。144ビット幅チャンネル340には、128のデータビットと16のECCビットとが含まれる。144ビットデータバス340は、4つの別々の72ビットバスセグメントに分割される。各バスセグメントは、データに使用される8つのデータメモリモジュール330と、ECCに使用される2つのECCメモリモジュール334と、1つのホットスペアメモリモジュール338と、を有しているのが好ましい。ECCメモリモジュール334は、各チャンネル346a~346dにおいて16のデータモジュールと4つのECCモジュールとに渡ってデータストライピングが可能であるように設けられている。

【0019】データバス、すなわちデータメモリチャンネル340は、少なくとも1つのスイッチング手段320、好ましくは1:2FETに電気的に結合されており、メモリコントローラ314を介してCPUバス31

0に電気的に結合されている。1:2FETスイッチであるスイッチング手段320は、144ビットバスを2つの144ビットバスセグメントに分割する。72ビット幅DRAMの場合、2つの144ビットバスセグメントの各々が、2つの別々の72ビットバスチャンネルに分割される。72ビットメモリチャンネルの各々は、複数の72ビットメモリモジュール330、334、338に並列に接続されている。

【0020】図3に示す実施の形態において、各メモリチャンネルは、2ECC DIMMの幅および36DIMMの深さであり、各々18のDIMMを持つ2つのバスセグメント（1:2FETスイッチによって分離されている）を使用している。各メモリチャンネルは、8つのデータDIMMと、2つのECC DIMMと、1つのホットスペアDIMMと、を有している。従って、メモリコントローラチャンネルデータI/Oピンにおける最大ローディングは、9DIMMである。これにより、4つのバスセグメントの各々において、1GBのDIMMを用いれば32GB、または2GBのDIMMを用いれば64GBの最大容量およびホットスペアDIMMが可能となる。

【0021】図3に示すメモリ構成は、ストライピングを使用する場合、容易に拡大縮小することができない。図3に示すDIMMの構成および数は、ハイエンドシステムでは一般的であるが、データバスに接続されたDIMMモジュールの数は、データ記憶用のメモリモジュール1つとホットスペアメモリモジュール1つとであってよい。データ記憶用メモリモジュールが1つだけ使用される場合は、DIMMは、エラー訂正専用の少なくとも1つのメモリ装置またはメモリ装置のビットを有していなければならない。

【0022】図4は、好ましい実施の形態の自己回復並列メモリ構成の部分ブロック図を示す。図4に示す好ましい実施の形態は、図3に示す実施の形態を縮小したバージョンである。ここに示すメモリ構成は、データまたはECCのいずれかを格納するために使用できる複数のメモリモジュール330と、ホットスペアメモリモジュール338a、338bと、を有している。図3に示す実施の形態とは対照的に、専用のECC装置は無く、代わりに、メモリ装置におけるビットがデータまたはECCに対して専用化されている。図4に示す実施の形態では、好ましくは、各72ビットバス360aまたは360bに提供される単一のホットスペアメモリモジュールがある。メモリモジュール330の1つに障害が発生した場合、データはその対応するホットスペアモジュールにマップされる。図4に示す実施の形態では、データストライピングは使用されず、従って、システムは拡大縮小可能である。図4に示す実施の形態では、最低2つのデータメモリモジュールと2つのスペアメモリモジュールとが必要である。

【0023】図4に示す実施の形態では、データメモリチャンネル340は、メモリコントローラ314を、少なくとも1つのスイッチング手段320、好ましくはFETスイッチに電氣的に結合する。FETスイッチ320は、メモリバスからの信号を分割して、メモリコントローラにおけるローディングを低減する。好ましい実施の形態では、スイッチング手段は、1:2 FETスイッチである。しかしながら、代替的な実施の形態では、スイッチング手段は、クロスバタイプスイッチとすることができる。しかしながら、システム要件により、特にメモリ構成におけるメモリモジュールの数により、1:2または1:4 FETを使用することができる。FETのサイズの比率が増加するということは、バスセグメントの数が増加することを意味する。

【0024】図4に示す好ましい実施の形態では、各72ビットバス360aまたは360bに対し、FETスイッチが使用されている。代替的な実施の形態では、メモリコントローラとCPUバスとの間にスイッチング手段が接続されていない。一般に、わずかな数のメモリモジュールしか必要でない小さいメモリシステム（例えば、1GBレンジ）では、メモリ構成の一部としてスイッチング手段は含まれない。

【0025】好ましい実施の形態では、メモリモジュール324は、JEDEC準拠を満たすために、72ビット幅のDRAM、好ましくはDIMMである。しかしながら、システム要件を満たす限りは、いかなるタイプの読出し書込み可能メモリコンポーネントを使用してもよい。メモリモジュール324は、スベアであっても稼動中であってもよい。従って、バンク選択、RAS等の制御ラインは、可変であり、好ましくは各スロットに対してハードワイヤード（hardwired）されていないことが必要である。

【0026】好ましい実施の形態では、メモリ構成の各メモリチャンネルは、少なくとも1つのデータ訂正用メモリモジュールと、少なくとも1つのホットスベアとして使用されるメモリモジュールと、を有している。少なくとも1つのデータ訂正用メモリモジュールおよび少なくとも1つのホットスベアとして使用されるメモリモジュールは、データバスに並列に電氣的に結合されている。好ましくは、メモリモジュールはSDRAMであり、システムは、SDRAM障害全体について訂正を行うことができると共に、ホットスベアにおける再構築プロセス中に発生する可能性のある、追加的なソフトエラーについても訂正を行うことができる。また、制御またはアドレスが冗長な接点を有する場合は、DIMMソケット障害についても訂正を行うことができる。

【0027】メモリ構成は、各メモリチャンネルに対し、2つのECCメモリモジュール334を有している。しかしながら、ECCモジュールの好ましい数は、使用されるCRC多項式のタイプに限定されず、データエラー

の数およびデータエラーが訂正されるチャンネルの数を含み、複数の要素によって決まる。メモリコントローラ314は、メモリエラーに対しデータ訂正を提供するECCロジック342を含む。ECCロジック342の実現には従来からの技術を用い、それは、使用されるデータ訂正要件およびアルゴリズムによって変更可能である。

【0028】使用されるCRC多項式とECCメモリモジュールの数とは、メモリ構成のデータ訂正要件によって変化する。好ましいECCの実現では、SDRAM4つの障害によるエラーを訂正するために16のECCビットが必要であり、同時に再構築プロセス中に1つのランダムエラーが許容される。これにより、ビットエラー訂正の数は合計5つとなる。

【0029】図3に示す実施の形態において、各メモリチャンネル346a~346dは、ホットスベアメモリモジュール338を有している。メモリ部分にホットスベアが存在することを宣言するために、構成ユーティリティプログラムが使用される。DRAM障害全体がDIMM（デュアルインラインモジュール）で検出された場合、それは訂正されるが、その後、メモリコントローラ314がホットスベアDIMM上にデータを再構築する。データビットに対する冗長ビットの割合は、1:8であるが、システムにおける冗長DIMM対DIMMの総数の割合により低減される。ホットスベアDIMMは、少なくとも最大のDIMMと同程度に大きくなければならない。一般に、ホットスベアメモリモジュールは、他のデータメモリモジュールと同じメモリサイズおよびDRAM編成を有している。

【0030】システム動作中、特定のメモリ位置において十分なエラーが発生した場合、データは、ホットスベアモジュールのメモリ位置に移動される。これにより、不良のデータメモリモジュールか、または異なるデータメモリモジュールの不良のデータメモリ位置が使用不可能となった場合、不良のデータメモリモジュールからのデータは、ホットスベアメモリモジュールに転送される。データメモリモジュール全体またはデータメモリモジュールの大部分が不良となった場合、そのデータメモリモジュールは、最終的にスワップアウトされ、データの完全性が向上しているホットスベアメモリモジュールと交換される。

【0031】図5から図7は、本発明の自己回復メモリで起こるステップを表す、高レベルブロック図を示す。訂正可能なハードエラーを有するコンピュータシステムに対し障害耐力のレベルを復元するために実行されるステップは、メモリモジュールにおける訂正可能エラーを検出することと、訂正可能エラーが検出されたメモリモジュールからのデータをスベアメモリモジュールに移動して再構築することを含む。

【0032】図5は、メモリエラーが第1のメモリモジュールにおいて発生している、図3および図4に示す構

成を表す高レベルブロック図を示す。図5を参照すると、第1のメモリモジュール330aにおいて陰を付けた領域で表された4ビットエラーが示されている。4ビットエラーは、メモリモジュールを読み出す際に検出され複数アドレスで発生する、訂正可能なハードエラーまたは永続的なエラーである。また、図5に示すブロック図には、スベアモジュール338aが示されている。

【0033】図6は、図3および図4に示す構成に対しデータをスベアメモリモジュールに移動して再構築するステップを表す高レベルブロック図を示す。図6を参照すると、矢印380は、データメモリモジュール330aからスベアメモリモジュール338aへのデータの移動を表している。移動および再構築プロセス中、データは第1のデータメモリモジュール330aからスベアメモリモジュール338aに転送される。移動および再構築プロセス中は、メモリコントローラによる第1のデータメモリモジュール330aへの新規な書き込みはすべて、データメモリモジュール330aとスベアメモリモジュール338aとの両方に対して行われる。矢印382は、この二重書き込みプロセスを表す。移動および再構築プロセスが完了するまで、読出し動作は第1のデータメモリモジュール330aからのみである。示されている移動および再構築プロセス中は、アルファ粒子により起こるような、一時的な1ビットエラーを訂正することができる。

【0034】図7は、第1のメモリモジュールからスベアメモリモジュールへのデータの移動および再構築のステップ後の、図3および図4に示す構成を表す高レベルブロック図を示す。図7を参照すると、以前はホットスベアモジュールであったモジュール5は、データモジュールになっている。古い第1のモジュールは、今では第1のモジュールからのオリジナルデータに新たな書き込みを足したデータを有している。システムユーザにとって有用である場合には、第1のメモリモジュールは、置換されるべきであり、古い第1のメモリモジュール330aはスベアメモリモジュールとなる。

【0035】ホットスベアメモリモジュールにデータを移動するために、メモリコントローラにより、不良のメモリ位置とホットスベアのメモリマッピングが行われる。メモリコントローラのデータメモリモジュールへのマッピングを変更することにより（すなわち、不良のDRAMを有するDIMMを検出した後の、アドレスまたは制御のルーティングにより）、ホットスベアデータメモリモジュールまたはDIMMは、書き込みおよびメモリ更新のためのホットスベアメモリモジュールと対となるが、読出しは不良のデータメモリモジュールのみから行われるように、設定される。

【0036】次に、CPU、メモリコントローラまたは管理コントローラは、不良のデータメモリモジュールの内容をホットスベアデータモジュールにコピーする。こ

のプロセス中、不良のデータメモリモジュールへのシステム書き込みもすべて、ホットスベアメモリモジュールに書き込まれる。従って、コピーが行われると、不良のデータメモリモジュールのイメージに再構築プロセス中に発生したあらゆる書き込みを足したものを、ホットスベアは有している。なお、DIMMからDIMMへの転送を行うときは初期化をしないのに対して、すべてのロケーションをコピーする場合は、訂正されたデータを有するホットスベアデータメモリモジュールを初期化することに留意してほしい。

【0037】エラー訂正方式は、同時に発生する複数のエラーを訂正することができなければならない。このため、例えば、DIMM再構築プロセス中に単一のビットソフトエラーが発生した場合、不良のDIMMに対し、新たなホットスベアとしてマップアウトする（そしてスワップアウトされるまで不良としてマークする）ことができない。そして、新たなホットスベアDIMMがメモリ構成内にマップされることにより、不良のDIMMが完全に置換される。スワップされるべきまたはエラーのログイングのためのDIMMを識別する際、実際の物理的なDIMM位置を計算するためにDIMMマップ機能が使用される。

【0038】例えば、図3において、システムがまだ実行中である時に、不良のデータメモリモジュールをホットスワップアウトするかまたは取出すようにしてもよい。データ転送が行われた後、新たなデータメモリモジュールにより不良のDIMMを置換することができるように、電源および不良DIMMを接続するバス（または取除くべきDIMM）がオフにされる。ホットスワッピングが好ましいが、サーバによっては、パッケージングは非常にアクセス密度が高いため、ホットスワッピングが提供されないかまたは実行不可能である場合がある。これらの場合、次に予定されたダウンタイム中に、不良のデータメモリモジュールをメモリ構成からコールドスワップすることができる。

【0039】好ましい実施の形態では、本発明の実現のために別々のスベアメモリモジュールが必要とされる。スベアメモリモジュールは、メモリ構成においてサポートされなければならない。一般的に、このスベアメモリサポートは、ホットスベアメモリモジュールによって受信されるコマンドおよびアドレスバスが、データモジュールおよびECCメモリモジュールに送信される信号とは異なっていることが必要である。これは、一般に、データモジュールまたはECCメモリモジュールにおいてエラーが発生しない限り、ホットスベアメモリモジュールは選択されないためである。訂正可能なハードエラーが発生する場合にのみ、ホットスベアメモリモジュールは起動されることになり、訂正不可能なエラーを有するメモリモジュールは休止状態になる。さらに、このスベアメモリモジュールサポートには、好ましくはメモリコ

ントローラにおいて実現されるか、または代替的にメモリコントローラ外部の論理回路において実現される、追加ロジックが含まれる。

【0040】再構築プロセスに対して最も柔軟な方式は、メモリコントローラが、障害の発生したDIMMを読み出し、同時にデータのバーストブロックを（メモリコントローラを経由して）ホットスเปアDIMMに転送する、というものである。障害の発生したDIMMとホットスเปアDIMMとは、同じメモリサイズおよび編成であるのが好ましいが、障害の発生したDIMMとホットスเปアDIMMとは、メモリサイズが異なってもよく、異なるSDRAM編成を使用してもよい。再構築プロセス中、システム読み出しは障害が発生したDIMMから行われ、システム書き込みは、別々のバーストサイクルで、障害が発生したDIMMとホットスเปアDIMMとの両方に対して行われる。

【0041】本発明の1つの実施の形態においては、メモリモジュールに対してデータを書込むために、ジャグラー (juggler) アルゴリズムが使用される。「ジャグラー」アルゴリズムは、同じ棍棒のセットを交互に空中で巧みに操る2人の曲芸師 (juggler) に例えてそう名づけられている。DIMMデータは、正常なシステムメモリのトランザクションが継続している間に、「オンラインで」不良DIMMからホットスเปアDIMMに徐々に移動されるかまたは交互に行われる (juggle)。好ましい実施の形態について、図4に示すメモリ書き込みパターンの各ブロックは、8ビットのデータを表しているが、ビットの数は実際のメモリ構成によって変えることができる。例えば、本発明の図3に示す構成における8ビットDIMMを16ビットDIMMモジュールで置換する場合、図4に示す各ブロックは、16ビットのデータを表すことになる。

【0042】DIMMにおけるDDR DRAMチップの破局故障か、またはDIMMコネクタ自体（データピン）の障害の後、システムは、サーバをシャットダウンする必要なく、自身をオンラインで修復し障害耐力の正常レベルに戻る。ジャグラーアルゴリズムを用いることで、本発明のメモリ構成は、以下のサーバアプリケーションに適する。すなわち、1) オペレータの居ない遠隔サーバロケーション、2) オンラインでホットスワッピングを行うにはアクセス不可能な程、非常に密度高くパッケージされたメモリシステムを有するサーバ、3) 何百または何千のサーバを含むことができる大規模集中型「サーバファーム」において使用されるよう設計され、I/O、CPUおよびメモリサブシステムのホットスワッピングに対し最適化されたサーバ、4) 障害のあるサーバを修復のためにシャットダウンすることができ、それゆえ個々のサブシステムにホットスワッピングを提供するという複雑さを要しない、冗長なサーバのペア、である。

【0043】本発明のメモリ構成において、JEDEC DDR DIMM仕様に準拠するように、 $\times 4$ および $\times 8$  SDRAMに対し133MHzの周波数でエラー訂正が動作する必要がある。DDRを100MHzで使用するデュアルチャネルのバースト帯域幅は、3.2GB/秒であり、133MHzの場合は4.2GB/秒である。従って、サーバまたはワークステーションのメモリシステムが、3.2GB/秒のバースト帯域幅で3:4の割合で100MHz  $4 \times$  CPUバスと共に使用される場合、非キャッシュコヒーレント  $4 \times$  AGPメモリアクセスに対し、ワークステーションで追加の帯域幅を使用することができる。あるいはまた、ホットスเปアDIMM再構築プロセス中に、キャッシュライン転送、I/Oバースト転送、またはメモリコントローラのオーバヘッドに要する、全待ち時間を短縮するため、サーバに帯域幅を追加して使用することができる。CPUバスが133MHzで動作する場合、メモリは166MHzで動作することができる。DDRバースト毎に転送されるデータ（バースト長が4とする）は、64バイトのキャッシュラインまたはCPUバスバーストに対応する。

【0044】好ましい実施の形態では、検出されたエラーは訂正され記録（ログ）される。イベントログはすべてのエラーを記録し、エラーログは異なるエラークラスに対しておよび特定のDIMMロケーションに対して行われるのが好ましい。なお、ソフトエラーは一時的な障害であり、ハードエラーは、永続的な障害を表す。このため、ニブル境界における4ビットハードエラーの検出により、ホットスเปア上のデータの「再構築」が起動される。同様に、DIMMコネクタ障害が検出された場合、優良なDIMMソケットを有するホットスเปア上のデータの再構築を起動するためにそれを使用することができ、あるいは、データを単に訂正することができる。

【0045】スパイラルな書き込みパターンを書込むためには、図3の構成に示すデータメモリモジュールはすべて存在しなければならない。メモリコントローラが異なるメモリアドレスを通過するに従い、データがスパイラルな形式または書き込みパターンでデータモジュールおよびECCメモリモジュールに書込まれる。ECCとラベル付けされているが、スパイラル書き込みパターンの場合、ECCモジュールはデータ用にも使用され、データモジュールは書き込みサイクルの場所によりECCデータの書き込みにも使用される。

【0046】複雑な方法ではあるが、スパイラル書き込みパターンより単純な再構築アルゴリズムを有する代替的な自己回復方法は、非ECC DIMMを使用し、ECC用に付加された2 DIMMを含む16バイトブロックのメモリチャネルに渡ってデータをストライプするというものである。それぞれ2つのホットスเปアDIMMが含まれる場合、これには最低20のDIMMが必要である。従って、データビットに対する冗長データビットの

割合は、ホットスワップを含んで1:4である。RAIDシステムと同様、DIMMはすべて同じサイズでなければならない。

【0047】データがアレイの各DIMMにストライプされると、ブロックサイズが1ニブルかまたは1バイト幅である場合、障害が発生したDIMMを、障害が発生した時に取り除き置換することができ、また、システムをクラッシュさせることなくDIMMを取除くことができる。その結果、不良DIMMのホットスワップ後、新たなDIMMは有効なデータを含まなくなり、メモリアクセスが行われるときに再構築することができ、結果として生じるニブルエラーに対する訂正が行われる。あるいはまた、全アドレス空間を、CPU、管理コントローラまたはメモリコントローラによってシーケンシャルに読出すことができる。これにより、再構築プロセスが行われるまで、データと共に書込まれる正しいECCビットを有する新たなDIMMの各ロケーションに対し、R-M-W (Read-Modify-Write) が行われる。

【0048】図8は、本発明のメモリモジュールに対してデータを書込むかまたはストライプする第1の方法を示す。図9は、本発明に従ってメモリモジュールに対してデータを書込むかまたはストライプする第2の方法を示す。図8に示す書込みパターンは72ビット幅のメモリモジュールに適用され、図3に示す実施の形態では、メモリチャンネル毎に少なくとも9つのDIMMと1つのスベアメモリモジュールが必要となる。対照的に、図9に示す書込みパターンは、64ビット幅のメモリモジュールに適用される。それは、一般的に、図8に示す方法と同じ数のDIMMが必要であるが、データメモリの容量が少ないため、ストライピングパターンが図8のパターンとは異なっている。

【0049】データストライピングという代替方法は、(1)ホットスワップ再構築アルゴリズムの複雑さを低減する(不良DIMMをいつでもスワップすることができる)こと、および(2)メモリコントローラとNOSが独立していることが魅力であるが、好ましくないかまたはシステム構成の柔軟性を低下させる可能性のある制約がいくつか追加される。第1に、メモリアレイは、同じサイズのDIMM内で完全に占有されなければならない。第2に、必要なDIMMの数は、4または18に分割されるDIMMデータ幅に等しい。これは、ベースシステムにおけるアレイサイズが非常に大きい、メモリアレイサイズを8ビットチップキルを用いることによって9トータルに縮小することができる、ということの意味する。この代替的な自己回復方式の問題は、データストライピングプロセスがメモリ制御およびアドレス指定の複雑さを増大させるのと同様に電力の消失を増大させる可能性がある、ということである。更に、OLXによる追加にホットスワップ方式が使用される場合、システムをパワーダウンしメモリアレイデータ全体を再

構築することなく、アレイに新たなDIMMを追加することが複雑になる。これら同じ問題は、RAIMと呼ぶこの方式に比較して、ディスクドライブを用いるRAIDシステムにも存在する。このため、この代替方法は、多くのDIMMソケットを備えた非常に大容量のメモリを有するハイエンドサーバにより適している。

【0050】図10は、RAMBUSメモリ構成と共に使用することができる本発明の第2の代替的な実施の形態の部分ブロック図を示す。図3に示す実施の形態に比べて、図10に示す代替的な実施の形態は好ましくない。図10に示すデージーチェーン直列アーキテクチャの基本的な問題は、システムの実行中には、データメモリモジュールをバスから取除くこともパワーダウンすることもできず、そのため、ホットスワッピングを行うことができない、ということである。更に、ECCビットは直列である。従って、1つのコネクタピン障害により、デージーチェーンにおける16のバースト的な複数エラーによりシステムに障害がもたらされることとなる。

【0051】図10を参照すると、メモリコントローラ412に電気的に接続されたCPUバス410が示されている。メモリコントローラ412は、複数のメモリチャンネル416a、416b、416cに電気的に接続されている。メモリチャンネル416aは、ホットスベアメモリモジュールを含んでいる。メモリチャンネル416bは、8つのデータメモリモジュールを含む。好ましくは、メモリチャンネル416cは、3つのECCメモリモジュールを含む。

【0052】図10に示すメモリ構成は、2バイト幅の非ECCメモリチャンネルを使用する。従って、RAMBUSモジュールの1チャンネル分を有するPCBを、システムから取り除くことができる。すなわち、8つのメモリチャンネル、2つのECCメモリチャンネルおよび1つのホットスベアチャンネルである。ECCビットの数は、2バイトのチャンネル障害を訂正するために十分大きくなければならない。データビットに対する冗長ビットの割合は小さい。従って、図10に示すメモリ構成は、非常にハイエンドなシステムに対してのみ適している。

【0053】上記説明は、例示のために示されており、限定することを意図しているのではないことが理解される。例えば、データメモリモジュールからのデータのマッピングを行うためにジャグアルゴリズムおよびデータストライピングアルゴリズムが示されているが、データをマッピングするために、RAIDメモリ構成において一般的に使用されるアルゴリズム等の、他のアルゴリズムを使用してもよい。従って、発明の範囲は、上記説明に関して決定されるべきではなく、添付の特許請求の範囲とかかる請求の範囲が権利を与えている同等物の全範囲に関して決定されるべきである。

【0054】本発明は例として次の実施態様を含む。



【0055】(1) メモリコントローラに電氣的に結合されたCPUバスと、メモリコントローラに電氣的に結合されたデータバスと、少なくとも1つのデータ記憶用メモリモジュールと、少なくとも1つのスเปアメモリモジュールと、を具備し、前記少なくとも1つのデータ記憶用メモリモジュールおよび前記少なくとも1つのスเปアメモリモジュールは、前記データバスに並列に電氣的に結合されており、前記少なくとも1つのデータ記憶用メモリモジュールで発生する訂正可能なエラーにตอบสนองして、前記少なくとも1つのデータ記憶用メモリモジュールからのデータが前記スเปアメモリモジュールにマップされるメモリ構成。

【0056】(2) 前記データ記憶用メモリモジュールは、複数のメモリ装置を含み、前記複数のメモリ装置の少なくとも一部は、エラー訂正に使用される、上記(1)に記載のメモリ構成。

【0057】(3) 前記少なくとも1つのデータ記憶用モジュールからのデータの前記スเปアメモリモジュールへの前記マッピングは、ユーザの介入無しに自動的に発生する、上記(1)に記載のメモリ構成。

【0058】(4) 少なくとも1つのエラー訂正用メモリモジュールを含み、前記少なくとも1つのエラー訂正用メモリモジュールは、前記少なくとも1つのデータ記憶用メモリモジュールおよび前記少なくとも1つのスเปアメモリモジュールに並列に電氣的に結合されている、上記(1)に記載のメモリ構成。

【0059】(5) 前記データバスを少なくとも第1のバスセグメントと第2のバスセグメントとに分割する少なくとも1つのスイッチを含む、上記(1)に記載のメモリ構成。

【0060】(6) 前記データバスを少なくとも第1のバスセグメントと第2のバスセグメントとに分割する前記少なくとも1つのスイッチがFETスイッチである、上記(1)に記載のメモリ構成。

【0061】(7) ジャグラルアルゴリズムを使用し

て、前記少なくとも1つのデータ記憶用メモリモジュールから前記スเปアメモリモジュールにデータがマップされる、上記(1)に記載のメモリ構成。

【0062】(8) データストライピングアルゴリズムを使用して、データがデータ記憶用メモリモジュールに書込まれる、上記(1)に記載のメモリ構成。

【図面の簡単な説明】

【図1】 DIMMモジュールが直列に接続されているRAMBUSメモリ構成の高レベル部分ブロック図。

【図2】 エラー訂正機能を提供するRAMBUSメモリ構成のデータおよびパリティチャンネルの高レベル部分ブロック図。

【図3】 本発明による自己回復並列メモリ構成の部分ブロック図。

【図4】 本発明の好ましい実施の形態による代替的な実施の形態の自己回復並列メモリ構成の部分ブロック図。

【図5】 第1のメモリモジュールでメモリエラーが発生している、図3および図4に示す構成を表す高レベルブロック図。

【図6】 図3および図4に示す構成についてスเปアメモリモジュールにデータを移動し再構築するステップを表す高レベルブロック図。

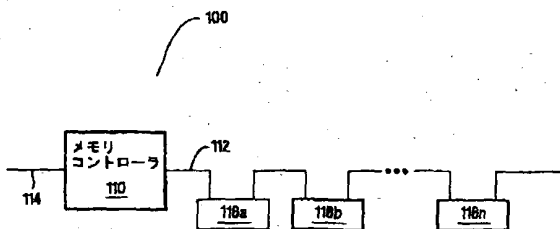
【図7】 第1のメモリモジュールからスเปアメモリモジュールへデータを移動し再構築するステップの後の、図3および図4に示す構成を表す高レベルブロック図。

【図8】 本発明による自己回復メモリ構成のメモリモジュールへのデータの書込みまたはストライピングの第1の方法。

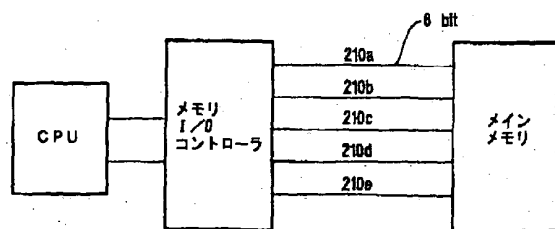
【図9】 本発明による自己回復メモリ構成のメモリモジュールへのデータの書込みまたはストライピングの第2の方法。

【図10】 RAMBUSメモリ構成と共に使用することができる本発明の代替的な実施の形態の部分ブロック図。

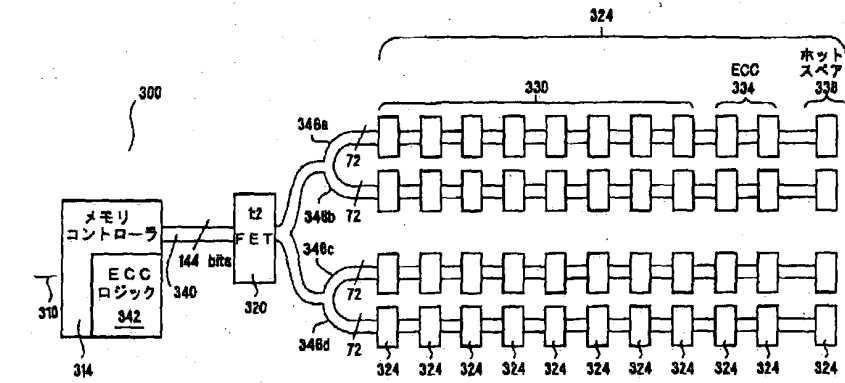
【図1】



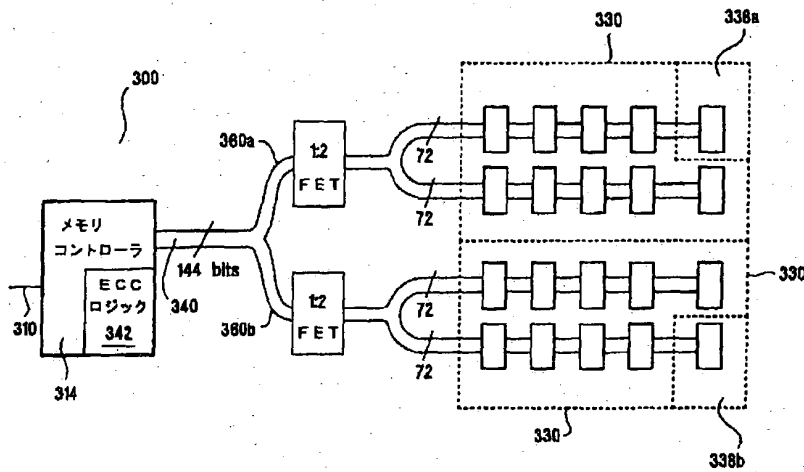
【図2】



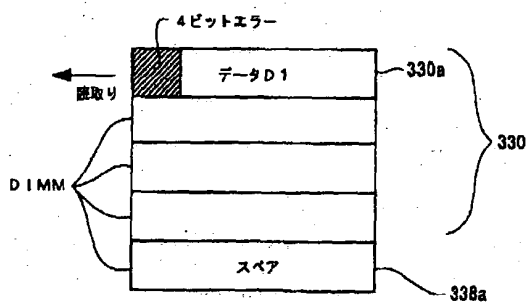
【図3】



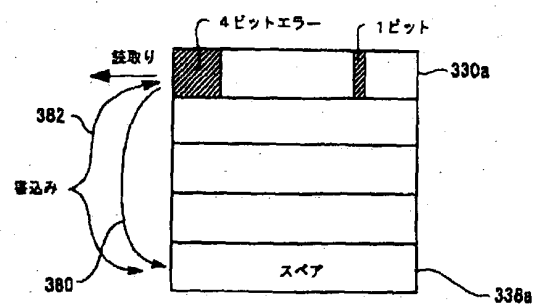
【図4】



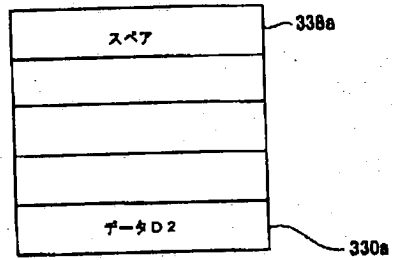
【図5】



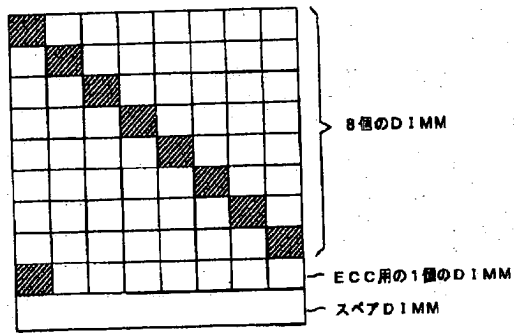
【図6】



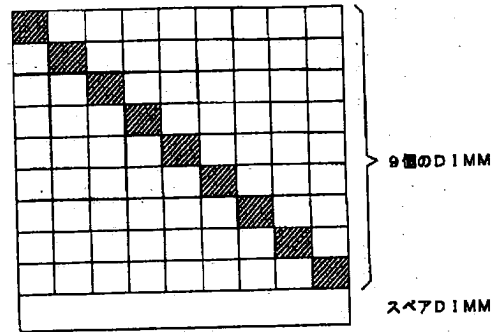
【図7】



【図9】



【図8】



【図10】

